WHAT IS CLAIMED IS:

1. A method for synthesizing speech, the method comprising:

generating a training context vector for each of a set of training speech units in a training speech corpus, each training context vector indicating the prosodic context of a training speech unit in the training speech corpus;

indexing a set of speech segments associated with a set of training speech units based on the context vectors for the training speech units;

generating an input context vector for each of a set of input speech units in an input text, each input context vector indicating the prosodic context of an input speech unit in the input text;

using the input context vectors to find a speech segment for each input speech unit; and

concatenating the found speech segments to form a synthesized speech signal.

2. The method of claim 1 wherein the each context vector comprises a position-in-phrase coordinate indicating the position of the speech unit in a phrase.

3. The method of claim 1 wherein the each context vector comprises a position-in-word

coordinate indicating the position of the speech unit in a word.

4. The method of claim 1 wherein the each context vector comprises a left phonetic coordinate indicating a category for the phoneme to the left of the speech unit.

5. The method of claim 1 wherein the each context vector comprises a right phonetic coordinate indicating a category for the phoneme to the right of the speech unit.

6. The method of claim 1 wherein the each context vector comprises a left tonal coordinate indicating a category for the tone of the speech unit to the left of the speech unit.

7. The method of claim 1 wherein the each context vector comprises a right tonal coordinate indicating a category for the tone of the speech unit to the right of the speech unit.

8. The method of claim 1 wherein the each context vector comprises a coordinate indicating a coupling degree of pitch, duration and/or energy with a neighboring unit.

9. The method of claim 1 the each context vector comprises a coordinate indicating a level of stress of a speech unit.

10. The method of claim 1 wherein indexing a set of speech segments comprises generating a decision tree based on the training context vectors.

11. The method of claim 10 wherein using the input context vectors to find a speech segment comprises searching the decision tree using the input context vector.

12. The method of claim 11 wherein searching the decision tree comprises:

identifying a leaf in the tree for each input context vector, each leaf comprising at least one candidate speech segments; and

selecting one candidate speech segment in each leaf node, wherein if there is more than one candidate speech segment on the node The selection is based on a cost function.

13. The method of claim 12 wherein the cost function comprises a distance between the input context vector and a training context vector associated with a speech segment.

14. The method of claim 13 wherein the cost function further comprises a smoothness cost that is based on a candidate speech segment of at least one neighboring speech unit.

15. The method of claim 14 wherein the smoothness cost gives preference to selecting a series of speech segments for a series of input context vectors if the series of speech segments occurred in series in the training speech corpus.

16. The method of claim 1 wherein the context vector comprises one or more higher order coordinates being combinations of at least two factors from a set of factors including:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit;

an indication of a category for a phoneme following a speech unit;

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit; and an indication of a degree of spectral mismatch with a neighboring speech unit.

17. A method of selecting sentences for reading into a training speech corpus used in speech synthesis, the method comprising:

identifying a set of prosodic context information for each of a set of speech units;

determining a frequency of occurrence for each distinct context vector that appears in a very large text corpus;

using the frequency of occurrence of the context vectors to identify a list of necessary context vectors; and

selecting sentences in the large text corpus for reading into the training speech corpus, each selected sentence containing at least one necessary context vector.

18. The method of claim 17 wherein identifying a collection of prosodic context information sets as necessary context information sets comprises:

determining the frequency of occurrence of each prosodic context information set across a very large text corpus; and

identifying a collection of prosodic context information sets as necessary context information sets based on their frequency of occurrence.

19. The method of claim 18 wherein identifying a collection of prosodic context information sets as necessary context information sets further comprises:

sorting the context information sets by their frequency of occurrence in decreasing order;

determining a threshold, F, for accumulative frequency of top context vectors; and

selecting the top context vectors whose accumulative frequency is not smaller than F for each speech unit as necessary prosodic context information sets.

20. The method of claim 17 further comprising indexing only those speech segments that are associated with sentences in the smaller training text and wherein indexing comprises indexing using a decision tree.

21. The method of claim 20 wherein indexing further comprises indexing the speech segments in the decision tree based on information in the context information sets.

22. The method of claim 21 wherein the decision tree comprises leaf nodes and at least one leaf node comprises at least two speech segments for the same speech unit.

23. A method of selecting speech segments for concatenative speech synthesis, the method comprising:

parsing an input text into speech units;

identifying context information for each speech unit based on its location in the input text and at least one neighboring speech unit;

identifying a set of candidate speech segments for each speech unit based on the context information; and

identifying a sequence of speech segments from the candidate speech segments based in part on a smoothness cost between the speech segments.

24. The method of claim 23 wherein identifying a set of candidate speech segments for a speech unit comprises applying the context information for a speech unit to a decision tree to identify a leaf node containing candidate speech segments for the speech unit.

25. The method of claim 24 wherein identifying a set of candidate speech segments further comprises pruning some speech segments from a leaf node based on differences between the context information of the speech unit from the input text and context information associated with the speech segments.

26. The method of claim 23 wherein identifying a sequence of speech segments comprises using a smoothness cost that is based on whether two neighboring candidate speech segments appeared next to each other in a training corpus.

27. The method of claim 23 wherein identifying a sequence of speech segments comprises using an objective measure comprising one or more first order components from a set of factors comprising:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit;

an indication of a category for a phoneme following a speech unit;

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit; and

an indication of a degree of spectral mismatch with a neighboring speech unit.

28. The method of claim 23 wherein identifying a sequence of speech segments comprises using an objective measure comprising one or more higher order components being combinations of at least two factors from a set of factors including:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit;

an indication of a category for a phoneme following a speech unit;

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit; and

an indication of a degree of spectral mismatch with a neighboring speech unit.

29. The method of claim 24 wherein identifying a sequence of speech segments further comprises identifying the sequence based in part on differences

between context information for the speech unit of the input text and context information associated with a candidate speech segment.

30. A computer-readable medium having computer executable instructions for synthesizing speech from speech segments based on speech units found in an input text, the speech being synthesized through a method comprising steps of:

identifying context information for each speech unit based on the prosodic structure of the input text;

identifying a set of candidate speech segments for each speech unit based on the context information;

identifying a sequence of speech segments from the candidate speech segments;

concatenating the sequence of speech segments without modifying the prosody of the speech segments to form the synthesized speech.